DEPARTAMENTO DE ECONOMÍA GENERAL Y ESTADÍSTICA

UNIDAD DOCENTE DE ESTADÍSTICA Y ECONOMETRÍA

TEMA 6

CORRELACIÓN

ASIGNATURA DE ESTADÍSTICA EMPRESARIAL

CURSO 2003-2004

FACULTAD DE CIENCIAS EMPRESARIALES

UNIVERSIDAD DE HUELVA

Primer Curso

Diplomatura en Ciencias Empresariales

PROFESORES:

David Castilla Espino

Encarnación Cordón Lagares

Concepción Cortés Rodríguez

Ramón Jiménez Toribio

Germán Pérez Morales

TEMA 6: CORRELACIÓN EN SPSS

Pág.

6.1. INTRODUCCIÓN 3

6.2. CORRELACIONES BIVARIADAS 3

6.3. CORRELACIONES PARCIALES 11

6.4. EJERCICIOS 14

6.5. BIBLIOGRAFÍA 16

6.1. INTRODUCCIÓN

Cuando se estudian dos variables (X,Y) o tres variables (X,Y,Z) es importante obtener una medida de la dependencia o medida de la relación entre esas variables.

Para estudiar y medir esta relación, el primer paso consistirá en recoger los datos que muestren los correspondientes valores de las variables consideradas y en representarlas después mediante un diagrama de dispersión o más sencillamente nube de puntos, como estudiamos en el tema 3. Esta representación gráfica es la que más se utiliza en el estudio de la dependencia de dos o tres variables, y son útiles como análisis previo a la ejecución de procedimientos de correlación y regresión.

Más tarde, estudiaremos la regresión entre dos variables que se refiere a hallar una fórmula o ecuación que represente la relación aproximada entre esas dos variables.

6.2.CORRELACIONES BIVARIADAS

El procedimiento "Correlaciones bivariadas de SPSS" permite medir el grado de dependencia existente entre dos o más variables mediante la cuantificación por los denominados coeficientes de correlación lineal de Pearson, de Spearman y la Tau-*b* de Kendall con sus respectivos niveles de significación.

Antes del cálculo de un coeficiente de correlación, inspeccionaremos los datos con el fin de detectar valores atípicos que puedan producir resultados equívocos.

Para la obtención de correlaciones bivariadas abriremos el archivo TTERRENO.SAV y procederemos como expresamos a continuación:

Elija en los menús: **Analizar ➔ Correlaciones ➔ Bivariadas**...

Una vez seleccionadas estas opciones, aparecerá el cuadro de diálogo que presentamos a continuación:

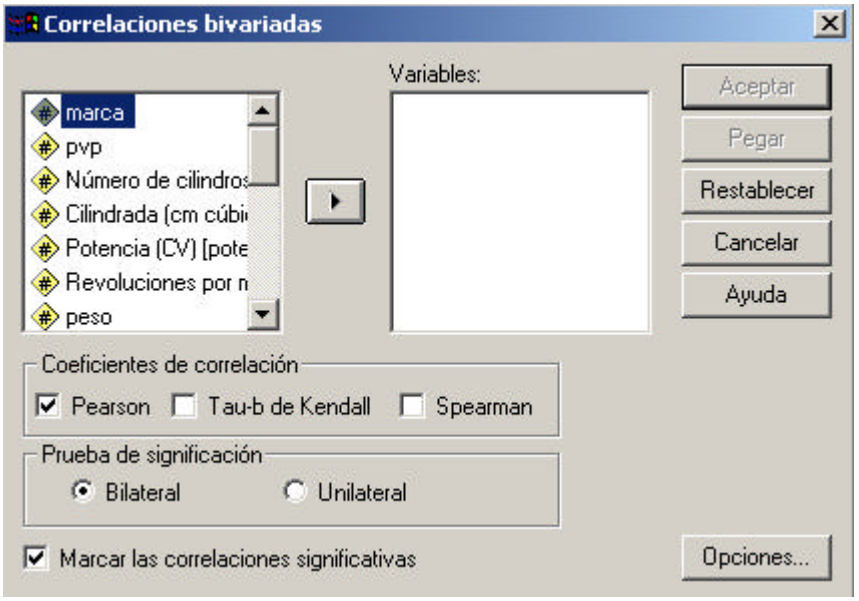


Tabla 1.- Cuadro: *Correlaciones bivariadas*

Introducimos en el recuadro de variables aquellas sobre las que vamos a cuantificar los coeficientes de correlación. Resulta obvio que al tratarse del cálculo de correlaciones, se deberán introducir al menos dos variables.

Ejemplo: (Archivo TTERRENO.SAV)

a) Calcula la relación existente entre las variables PVP y CONSURB.

b) Representa la dispersión simple de las variables anteriores.

c) Repite el apartado anterior estableciendo marcas según la variable CILINDRO y etiquetando los casos mediante la variable MODELO.

Una vez seleccionadas aquellas variables, cuyos coeficientes de correlación vamos a cuantificar, deberemos seleccionar el tipo de coeficiente de correlación que queremos calcular, para lo cual se nos presentan tres opciones diferentes:

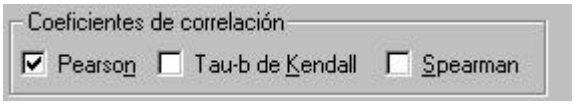


Tabla 2.- Correlaciones bivariadas: *Coeficientes de correlación*

- ***Pearson*** ➔ Es una medida de la asociación lineal entre dos variables. Es el más conocido y utilizado de todos. Toma valores que se encuentran dentro del intervalo cerrado [-1,1], pero un valor de -1 o +1 sólo se puede obtener a partir de tablas cuadradas. El signo del coeficiente indica la dirección de la relación siendo ésta directa, para el caso de valores positivos, e inversa, para el caso de valores negativos. Su valor absoluto indica la fuerza de la misma, de tal modo que, los mayores valores indican que la relación de dependencia entre las dos variables es más estrecha, en el sentido de más fuerte. Un valor de 0 indica o que las dos variables X e Y son independientes o que no existe una relación de tipo lineal entre ambas variables.

- ***Tau-b de Kendall*** ➔ Medida no paramétrica de asociación para variables ordinales o de rangos que tiene en consideración los empates. El signo del coeficiente indica la dirección de la relación y su valor absoluto indica la magnitud de la misma, de tal modo que los mayores valores absolutos indican relaciones más fuertes. Los valores posibles varían de -1 a 1, pero un valor de -1 o +1 sólo se puede obtener a partir de tablas cuadradas.

- ***Spearman*** ➔ Versión no paramétrica del coeficiente de correlación de Pearson, que se basa en los rangos de los datos en lugar de hacerlo en los valores reales. Resulta apropiada para datos ordinales (susceptibles de ser ordenados) y para datos agrupados en intervalos que no satisfagan el supuesto de normalidad. Los valores del coeficiente varían de -1 a +1. El signo del coeficiente indica la dirección de la relación y el valor absoluto del coeficiente de correlación indica la fuerza de la relación entre las variables. Los valores absolutos mayores indican que la relación es mayor.

NOTA:

Para las variables cuantitativas y normalmente distribuidas, seleccionaremos el coeficiente de correlación de Pearson, mientras que si los datos no están normalmente distribuidos o tienen categorías ordenadas, seleccionaremos la Tau-*b* de Kendall o el coeficiente de correlación por rangos de Spearman, que miden la asociación entre órdenes de rangos.

De entre estos coeficientes vamos a centrarnos en el coeficiente de correlación lineal de Pearson entre dos variables X e Y, el cual ha sido objeto de estudio en esta asignatura. Su expresión matemática es la que presentamos a continuación[1]:

$$r = \frac{S_{XY}}{S_X \cdot S_Y} = \frac{\sum_{i=1}^{h}\sum_{j=1}^{k}(x_i - \bar{x})\cdot(y_j - \bar{y})n_{ij} \Big/ (N-1)}{S_X \cdot S_Y} \qquad -1 \le r \le 1$$

Puede ocurrir que dos variables estén perfectamente relacionadas y que la relación no sea de tipo lineal. En este caso diremos que, el coeficiente de correlación de Pearson no es un estadístico adecuado para medir su grado de asociación.

Tras la especificación de los coeficientes de correlación que estimemos oportuno calcular, procederemos a indicar si queremos que se realice una prueba de significación o contraste de hipótesis de tipo bilateral (de dos colas) o de tipo unilateral[2] para casos en los que la dirección de la relación puede ser especificada a priori.



Tabla 3.- Correlaciones bivariadas: *Pruebas de significación*

Este contraste, trata de probar la hipótesis de que el coeficiente de correlación sea nulo (r =0), esto es, que no exista relación alguna entre las variables cuyo coeficiente de correlación estamos cuantificando. El contraste de hipótesis al que hacíamos referencia en el párrafo anterior utiliza como estadístico de contraste a *t*, el cual se distribuye según una función de probabilidad *t* - *S*tudent[3] con *N-2* grados de libertad siempre y cuando las variables cuya correlación se mide, se distribuyan normalmente.

[1] Obsérvese que en el tema 7 de teoría se utilizó el número total de observaciones, N, en lugar de N-1.
[2] Un contraste de hipótesis es un test que se utiliza para contrastar dos hipótesis.
[3] Una distribución *t* de *S*tudent se puede definir en términos de una distribución normal y una χ^2 independientes. Sean Z~N(0,1) y V~χ^2_r donde Z y V son ambas independientemente. Entonces, $t = \frac{z}{\sqrt{\frac{V}{r}}} \sim t_r$

$$t = r \cdot \sqrt{\frac{N-2}{1-r^2}} \rightarrow t_{N-2}$$

Finalmente, y con el objeto de identificar aquellos coeficientes de correlación que tienen una mayor significación, se puede seleccionar la opción:

Marcar las correlaciones significativas,

Dicha opción que marca los coeficientes de correlación significativos al nivel 0,05 (5%) por medio de un solo asterisco y los significativos al nivel 0,01 (1%) con dos.

Ejemplo: (Archivo TTERRENO.SAV)

a) Calcula la relación existente entre las variables PVP, CILINDRO, CC y POTENCIA.

b) Representa la dispersión matricial entre las variables PVP, CC y POTENCIA.

La salida que SPSS proporciona es la que presentamos a continuación:

Correlaciones

		PVP	Número de cilindros	Cilindrada (cm cúbicos)	Potencia (CV)
PVP	Correlación de Pearso	1,000	,641**	,696**	,731**
	Sig. (bilateral)	,	,000	,000	,000
	N	125	125	125	125
Número de cilindros	Correlación de Pearso	,641**	1,000	,703**	,730**
	Sig. (bilateral)	,000	,	,000	,000
	N	125	125	125	125
Cilindrada (cm cúbicos)	Correlación de Pearso	,696**	,703**	1,000	,752**
	Sig. (bilateral)	,000	,000	,	,000
	N	125	125	125	125
Potencia (CV)	Correlación de Pearso	,731**	,730**	,752**	1,000
	Sig. (bilateral)	,000	,000	,000	,
	N	125	125	125	125

**. La correlación es significativa al nivel 0,01 (bilateral).

Tabla 4.- Resultados obtenidos de la opción: *Correlaciones bivariadas*

Esta salida consiste en una matriz de tipo simétrica que toma valores unitarios en la diagonal. Se muestra el coeficiente de correlación seleccionado (*r*) para cada par de variables, la significación del contraste realizado (*p*), de modo que cuanto menor sea su valor más fiable será el dato arrojado por el coeficiente de correlación seleccionado, y el número de casos no perdidos considerados (*N*).

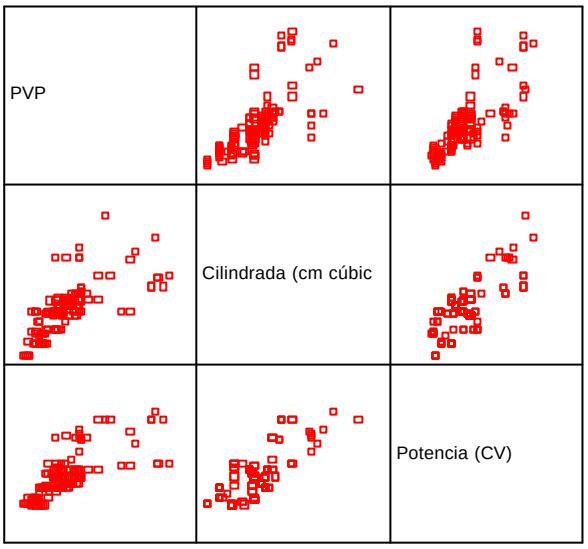


Figura 1.- Resultados obtenidos de la opción: *Gráficos /Dispersión /Matricial*

Si se ha seleccionado la opción ***Marcar las correlaciones significativas*** se podrá observar las marcas en las correlaciones que cumplan las condiciones anteriormente establecidas.

Si pulsamos en el botón ***Opciones*** del cuadro de diálogo, SPSS nos ofrece la posibilidad de calcular diversos estadísticos:

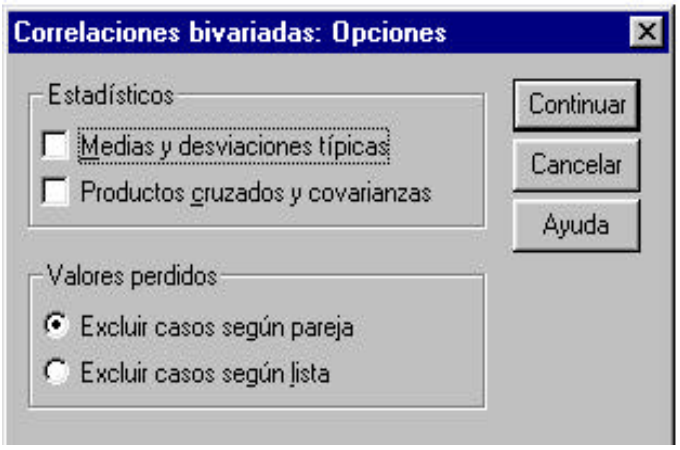


Tabla 5.- Correlaciones bivariadas: *Opciones*

- **Medias y desviaciones típicas** ➔ Si elegimos esta opción, SPSS calcula para cada una de las variables que se han introducido en el cuadro de variables la media y la desviación típica. También se muestra el número de casos que no tienen valores perdidos.

Si seleccionamos esta opción y ejecutamos el procedimiento, la salida que nos muestra SPSS es la que presentamos a continuación:

Estadísticos descriptivos

	Media	Desviación típica	N
PVP	4004459	1937790,98	125
Número de cilindros	4,59	1,02	125
Cilindrada (cm cúbicos)	2569,82	691,51	125
Potencia (CV)	117,08	37,57	125

Tabla 6.- Correlaciones bivariadas: *Estadísticos Descriptivos*

- **Productos cruzados y covarianzas** ➔ Si elegimos esta opción, SPSS calcula para cada par de variables el producto cruzado de las desviaciones, que es igual a la suma de los productos de las variables corregidas respecto a la media, esto es, el numerador del coeficiente de correlación de Pearson; y la covarianza que es una medida no tipificada de la relación entre dos variables, igual al producto cruzado diferencial dividido por *N-1*.

$$S_{XY} = \frac{\sum_{i=1}^{h}\sum_{j=1}^{k}\left(x_i - \overline{x}\right)\cdot\left(y_j - \overline{y}\right)n_{ij}}{N-1}$$

Si seleccionamos esta opción y ejecutamos el procedimiento, la salida que nos muestra *SPSS* es la que presentamos a continuación:

Correlaciones

		PVP	Número de cilindros	Cilindrada (cm cúbicos)	Potencia (CV)
PVP	Correlación de Pearson	1,000	,641**	,696**	,731**
	Sig. (bilateral)	,	,000	,000	,000
	Suma de cuadrados y productos cruzados	4,656E+14	156546736,78	1,1562E+11	6600662211
	Covarianza	3,755E+12	1262473,684	932382471,697	53231146,86
	N	125	125	125	125
Número de cilindros	Correlación de Pearson	,641**	1,000	,703**	,730**
	Sig. (bilateral)	,000	,	,000	,000
	Suma de cuadrados y productos cruzados	156546737	128,192	61247,616	3460,080
	Covarianza	1262473,7	1,034	493,932	27,904
	N	125	125	125	125
Cilindrada (cm cúbicos)	Correlación de Pearson	,696**	,703**	1,000	,752**
	Sig. (bilateral)	,000	,000	,	,000
	Suma de cuadrados y productos cruzados	1,156E+11	61247,616	59294268,768	2421548,840
	Covarianza	932382472	493,932	478179,587	19528,620
	N	125	125	125	125
Potencia (CV)	Correlación de Pearson	,731**	,730**	,752**	1,000
	Sig. (bilateral)	,000	,000	,000	,
	Suma de cuadrados y productos cruzados	6,601E+09	3460,080	2421548,840	175027,200
	Covarianza	53231147	27,904	19528,620	1411,510
	N	125	125	125	125

**. La correlación es significativa al nivel 0,01 (bilateral).

Tabla 7.- Resultados de Correlaciones bivariadas: *Opciones*

De otra parte, en relación con los valores perdidos se pueden seleccionar las siguientes opciones:

- **Excluir casos según pareja** ➔ Se excluyen del análisis los casos (o filas) con valores perdidos para una o ambas variables de una pareja (X,Y) que forma un coeficiente de correlación. Debido a que cada coeficiente está basado en todos los casos que tienen códigos válidos para esa pareja concreta de variables, en cada cálculo se utiliza la mayor cantidad de información disponible. Esto puede dar como resultado un grupo de coeficientes basados en un número de casos variables.

- **Excluir casos según lista** ➔ Se excluyen de todas las correlaciones los casos con valores perdidos para cualquieras de las variables de la lista.

6.3.CORRELACIONES PARCIALES

Este tipo de coeficientes de correlación describe la relación lineal existente entre dos variables sin tener en cuenta los efectos o influencias de una o más variables adicionales, con el objeto, bien de identificar la existencia de posibles variables interpuestas, o de correlaciones neutralizadas por el efecto de estas variables.

Por tanto, puede ocurrir que dos variables estén perfectamente relacionadas pero si la relación entre ellas no es lineal, entonces el coeficiente de correlación no será un estadístico adecuado para medir su asociación.

Para la obtención de correlaciones parciales en SPSS, procederemos como sigue a continuación:

Elija en los menús: **Analizar ➔ Correlaciones ➔ Parciales**...

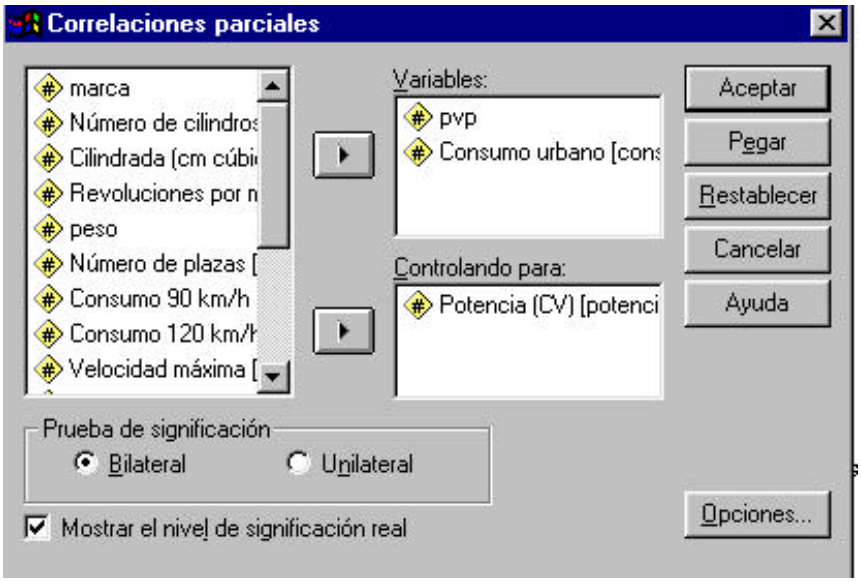


Tabla 8.- Correlaciones parciales

Una vez seleccionadas estas opciones, aparecerá el cuadro de diálogo que presentamos a continuación:

Al igual que ocurría con el cuadro de diálogo del comando *Correlaciones bivariadas* el cuadro *Correlaciones Parciales* presenta un formato similar al del resto de cuadros de diálogo que posee el paquete SPSS, de modo que lo que tendremos que hacer en primera instancia, es introducir en el recuadro de variables elegidas para el análisis aquellas sobre las que vamos a cuantificar los coeficientes de correlación parcial. Resulta obvio que al tratarse del cálculo de correlaciones, se deberán introducir al menos dos variables.

Una vez seleccionadas aquellas variables cuyos coeficientes de correlación vamos a cuantificar, deberemos introducir aquella variable o variables de control que estimemos conveniente, con el objeto de eliminar su efecto en la correlación de las variables introducidas en el recuadro de variables.

Ejemplo: (Archivo TTERRENO.SAV)

Calcula la relación existente entre las variables PVP y CONSURB utilizando como variable de control la POTENCIA.

Al igual que con el procedimiento de correlaciones bivariadas, para el caso de correlaciones parciales, también es posible realizar un contraste de hipótesis estadística bilateral o unilateral para casos en los que la dirección de la relación puede ser especificada a priori.



Tabla 9.- Prueba de significación de la opción: *Correlaciones parciales*

Este contraste, trata de probar la hipótesis de que el coeficiente de correlación sea nulo, esto es, que no exista relación alguna entre las variables cuyo coeficiente de correlación estamos cuantificando; no obstante, a diferencia del caso de correlación bivariada, el estadístico de contraste usado en correlación parcial es el que expresamos a continuación:

$$t = r \cdot \sqrt{\frac{N-q-2}{1-r^2}} \rightarrow t_{N-q-2}$$

Este estadístico se distribuye según una función de probabilidad *t*-Student con $N-q-2$ grados de libertad, donde q es el orden del coeficiente de correlación parcial, definido por el número de variables de control. Suponemos también la distribución normal de las variables objeto de análisis, *N* el número de casos y *r* el coeficiente de correlación parcial.

Finalmente, si seleccionamos la opción ***Nivel de significación real***, se muestran la probabilidad y los grados de libertad para cada coeficiente en la salida para este procedimiento.

La salida que SPSS proporciona es la que presentamos a continuación:

Corr. parciales

```
- - -  P A R T I A L   C O R R E L A T I O N   C O E F F I C I E N T S  -
- -

Controlling for..    POTENCIA

                  PVP     CONSURB

PVP            1.0000      -.1647
              (    0)     (  115)
              P= .        P= .076

CONSURB        -.1647      1.0000
              (  115)     (    0)
              P= .076     P= .

(Coefficient / (D.F.) / 2-tailed Significance)

" . " is printed if a coefficient cannot be computed
```

Tabla 10.- Resultados de la opción: *Correlaciones parciales*

La salida que hemos mostrado pone de manifiesto que frente a un coeficiente de correlación de Pearson para las variables PVP y CONSURB de 0.5140 hemos pasado tras la eliminación del efecto de la variable interpuesta POTENCIA a una correlación de –0.1647, lo que nos permite concluir que la potencia de los coches es la variable que determina que entre las variables PVP y CONSURB exista una relación positiva y significativa.

Si pulsamos en el botón ***Opciones*** del cuadro de diálogo, SPSS nos ofrece la posibilidad de calcular los mismos estadísticos que el procedimiento de correlaciones bivariadas, además de las correlaciones de orden $\boldsymbol{q} = 0$, es decir, de los coeficientes de correlación ordinarios, sin variable de control.

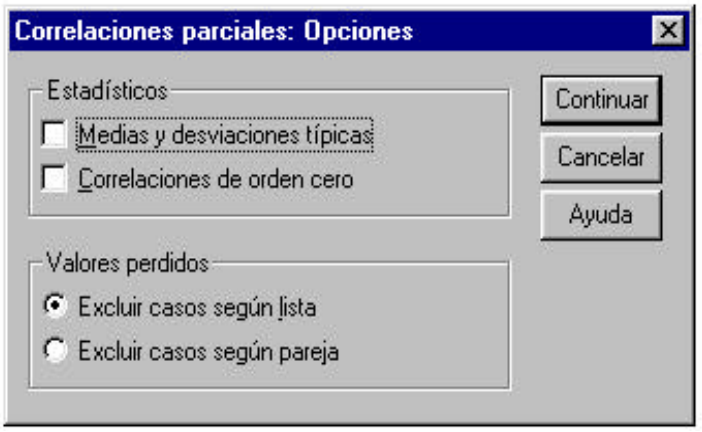


Tabla 11.- Correlaciones parciales: *Opciones*

6.4.EJERCICIOS

6.4.1. Utilizando el archivo TTERRENO.sav, se pide:

a) Calcula la relación entre las variables que representan las CILINDRADAS (en cm cúbicos), la POTENCIA (en CV) y el CONSUMO a 120 km/h, a través del coeficiente de correlación lineal.

b) Calcula la relación entre las variables que representan las CILINDRADAS (en cm cúbicos) y la POTENCIA (en CV) controlado por la variable CONSUMO a 120 km.

c) Interpreta el coeficiente de correlación lineal en ambos casos.

d) Representa gráficamente la dispersión matricial de las 3 variables anteriores.

6.4.2. Utilizando el archivo TRABAJO.sav vamos a buscar la posible relación entre estas tres variables:

b1: Situación laboral

c1: Sexo

c3: Estado civil

Completar la siguiente tabla para las tres variables utilizando primero la opción en relación con los valores perdidos de excluir casos según pareja y posteriormente excluir casos según lista.

VARIABLE	SEGÚN PAREJA			SEGÚN LISTA		
	SUJETOS	MEDIA	DESV. TÍP.	SUJETOS	MEDIA	DESV. TÍP.
b1						
c1						
c3						

Observa las diferencias tanto en número de individuos como en el valor de algunos estadísticos a que ambas opciones pueden dar lugar y completa el siguiente cuadro para el cual tendrás que calcular la matriz de coeficientes de correlación de Pearson también utilizando ambas opciones en relación con los valores perdidos:

VARIABLES	ESTADÍSTICOS	SEGÚN PAREJA	SEGÚN LISTA
b1 – c1	*r* de Pearson		
	N		
	p grado de significación		
b1 – c3	*r* de Pearson		
	N		
	p grado de significación		
c1 – c3	*r* de Pearson		
	N		
	p grado de significación		

Intenta buscar las diferencias entre ambas opciones utilizadas y dar una explicación a las mismas.

6.4.3. Repite el ejercicio anterior con las tres variables siguientes:

c5: Números de miembros del hogar

c6: Estudios del entrevistado

c9: Ideología política

6.4.4. Utilizando el archivo TRABAJO.SAV vamos a llevar a cabo un estudio de relación entre las variables c2 y c17 (EDAD y NIVEL DE INGRESOS MENSUALES DEL HOGAR) para toda la muestra, obteniendo un coeficiente de correlación lineal de_____. Si ajustamos esta relación a la variable c6 (ESTUDIOS DEL ENTREVISTADO) observamos que el coeficiente de correlación parcial disminuye a _____, con un grado de significación todavía de _____. Una vez calculados ambos coeficientes interpreta los resultados obtenidos.

6.5.BIBLIOGRAFÍA

CAMACHO ROSALES, J., (2002), *Estadística con SPSS para Windows (versión 11)*, Rama, Madrid.

MARTÍN PLIEGO, F.J., (1994), *Introducción a la Estadística Económica y Empresarial (Teoría y Práctica)*, AC, Madrid.

PALMER POL, A.L., (1999*), Análisis de Datos. Etapa Exploratoria, Psicología*, Pirámide, Madrid.

PÉREZ LÓPEZ, C., (2002), *Estadística aplicada a través de Excel*, Prentice Hall, Madrid.

SPSS INC., (1999*), Manual del Usuario de SPSS Base 10.0*, SPSS, Chicago.

VISAUTA VINACUA, B., (2002), *Análisis Estadístico con SPSS para Windows, volumen I, Estadística Básica, 2ª Edición*, McGraw-Hill, Madrid.